

DOCUMENT RESUME

ED 243 962

TM 840 283

AUTHOR Constable, Elizabeth; Andrich, David
TITLE Inter-Judge Reliability: Is Complete Agreement among Judges the Ideal?
PUB DATE Apr 84
NCTE 22p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 24-26, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary Secondary Education; Evaluation Methods; *Interrater Reliability; Interviews; *Item Analysis; Judges; Latent Trait Theory; Measurement Techniques; *Models; *Rating Scales; Social Cognition; Testing Problems
IDENTIFIERS *Attenuation Paradox

ABSTRACT

In circumstances where judges are required to make ratings of performance, it is usually required to have two or more raters who are trained to agree on independent ratings of the same performance. It is suggested that such a requirement may produce the paradox of attenuation associated with item analysis, in which too high a correlation between items, while enhancing reliability, decreases validity. With an empirical example of ratings of levels of social understanding it is shown how information is lost as judges agree more and more over a series of ratings which follow discussions of the previous set of ratings. Pursuing the item analysis analogy, it is demonstrated how it might be more appropriate to handle systematic differences among judges using a person-item response theory approach. (Author)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

INTER-JUDGE RELIABILITY: IS COMPLETE
AGREEMENT AMONG JUDGES THE IDEAL?

Elizabeth Constable

and

David Andrich

Department of Education
University of Western Australia
Nedlands, Western Australia 6009

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as
received from the person or organization
originating it.

☐ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official NIE
position or policy.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

D. Andrich

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

ABSTRACT

In circumstances where judges are required to make ratings of performance, it is usually required to have two or more raters who are trained to agree on independent ratings of the same performance. It is suggested that such a requirement may produce the paradox of attenuation associated with item analysis, in which too high a correlation between items, while enhancing reliability, decreases validity. With an empirical example of ratings of levels of social understanding it is shown how information is lost as judges agree more and more over a series of ratings which follow discussions of the previous set of ratings. Pursuing the item analysis analogy, it is demonstrated how it might be more appropriate to handle systematic differences among judges using a person-item response theory approach.

Presented at the Conference of the National Council of Measurement in
Education, New Orleans, 1984.

1. INTRODUCTION

Social scientists who gather empirical data know when they need to obtain inter-observer agreement, the difficulties associated with obtaining it, and the procedures required for enhancing it. For example, within education, inter-observer (grader) agreement is considered essential, not only in research, but also in practice. Quite generally, it is required that students' grades do not depend on which grader, from an appropriate class of graders, grades any particular performance.

The tasks of observers usually are one of two related types. First, responses, performances or behaviours are classified into mutually exclusive categories according to some specified system. Observations of classroom interactions are of this kind. Second, and in addition to a classification, these responses are ordered or rated. Effectively, the classification system has ordered categories. Systems for grading essays are often of this kind. Dawes (1972) points out that rating methods are extremely common in the social sciences with some 60 per cent of studies having dependent variables of this form.

Ratings by observers or expert judges are required usually where there is no direct procedure or instrument for obtaining measurements. Methods for considering the effects of graders in data collection designs have been well documented from the perspective of traditional test theory (Guilford, 1954; Stanley; 1961). The multiple-choice testing movement was motivated essentially by the desire to remove completely any effects due to the graders. Indeed, because the score of a person was independent of any grader, for a long time such tests were termed "objective". Despite more than a half-century of research, multiple-choice tests have not displaced completely the need for expert observers in education and psychology who can rate performances. For example, it is considered that certain complex performances would be rendered invalid if they were transformed into a multiple-choice

format, even if the performances on the multiple-choice tests are correlated highly with the original performances. Such arguments are put forward with respect to the assessment of essay-writing ability. Ratings are also deemed necessary in research and applied studies where clinical interviews are used in order to understand the structure of the thought processes of an individual. Such circumstances arise when persons are rated into levels formulated according to a stage theory of development (Keating and Clark, 1980). The example described later in the paper is of this kind.

This paper re-examines the rationale behind the desire and general requirement to have full inter-observer agreement. It will then be suggested that this requirement is not only unnecessarily restrictive, but that followed blindly, it may actually reduce the validity of the observations. That the topic is even being considered, let alone the position being advanced, may seem surprising. This position also came as a surprise to the authors when they attempted to understand a set of empirical data in which it was possible to follow one aspect of the process of obtaining inter-observer agreement. Thus it is a set of empirical data, the details of which are provided later in the paper, which have provoked this apparently unusual position.

The first part of the argument, that inter-observer agreement may be unnecessarily restrictive, has been studied and discussed earlier by Andrich and Hake (1979). In that work, concerned with the judgments of essay-writing ability, it was concluded that to have graders conformable in their observations to the point where the observations of one grader could be transformed into those of another was sufficient to satisfy the requirements of inter-grader reliability. It is the second part of the case, that obtaining total agreement at all costs may not be the ideal, which we wish to pursue here.

2. THE ROLE OF THE OBSERVER

Studies in relation to inter-observer agreement, and the process by which this agreement is reached, has forced us to reconsider the role of the observer and to re-examine the most primitive or fundamental aspects of agreement between observers. In doing so, we cannot avoid the position of Kuhn (1970) which stresses the role of the researcher in the construction of scientific knowledge. It has perhaps become somewhat *passé* to invoke Kuhn in order to explain or understand any apparent incongruities in the scientific process. However, his documentation of the importance of the researcher's psychological make-up in relation to the interpretation of the data, together with the social and intellectual milieu in this process, is particularly compelling, even though it is distinctive and is the source of a great deal of controversy within the philosophy and sociology of science. His position is so easily understood by active empirical researches (Masterman, 1970), that they cannot help but empathise with his description of the conduct of science.

Replication

While many different aspects of Kuhn's work could be elaborated upon, it is the issue of replication of any observed relationships which we wish to highlight here. In doing so, we draw on Barnes (1982), who elaborates on the same topic in his sympathetic yet critical examination of Kuhn's work for the social sciences.

Replicability, the ability to replicate relevant observations, is essential to both the notions of reliability and validity. Replicability in what are seen as identical contextual circumstances is usually referred to as reliability. Thus checking the concordance in scores of the same persons on the same test may be seen as a check on the test's reliability. Validity, on the other hand, depends on replicability in circumstances which in some relevant sense are different. For example, checking the concordance in scores

of the same persons on two different tests purporting to test the same ability, may be seen as a check on validity.

To the degree that any relationship can be replicated in circumstances which in some relevant sense are different, to that degree the relationship is general, or as it may be said, valid.

Barnes comments on the issue of replicability, using efforts to detect gravity waves as an illustration, as follows:

Those who undertook to replicate this initial work proceeded to build their own apparatus for gravity-wave detection. Such apparatus was not, however, conceived as an indistinguishable replica of the original. Any particular apparatus was considered by its designers to be 'the same' as the original, to the extent that both were describable as 'gravity-wave detectors'; but it was also considered to be different...(Barnes, 1982: 88).

Further, he reiterates the point:

The very idea of replication seemed to involve the notion of recognised variation from the initial work (p.89).

Rasch, better known for his work in psychometric models, expresses the same idea in his quest for understanding the essentials of the concept of measurement:

If a relationship between two or more statistical variables is to be considered really important, as more than an ad hoc description of a very limited set of data - if a more or less general inter-dependence may be considered in force - the relationship should be found in several sets of data which differ materially in some relevant respects (Rasch, 1980: 9).

The Instrument

In the above examples, there is an implied distinction between the observer and the instrument used to generate the relevant events. However, the instrument is produced by someone in order to produce certain events which are to be observed and to which the observer is sensitised. Thus, in making the distinction between an instrument and the observer, we cannot dismiss the

role of the observer as immaterial. Whenever any interpretation of data is involved, and all observations are interpreted in order that meaning be given to them, the role of the observer is crucial. It seems that it is this aspect of Kuhn's description of science which creates great concern to those who wish to see science, and the body of scientific knowledge, as independent of the scientists who develop that knowledge.

Interpreting Observations

Of particular relevance for the case to be made here are the interpretations of different observers of events generated by the same instrument. It is here that total agreement is considered desirable. Yet here, too, the principle that perfect agreement may not constitute a genuine replication across observers, needs to be considered. To help make the point, suppose that identical twins were asked to interpret events generated by an instrument. If it were established that these twins performed and behaved identically in general, then obtaining the same results from the twins would not constitute a replication in observations which are different in any way. Agreement between identical twins would not be considered as strong evidence of inter-observer agreement.

Thus we suppose that the first stage in demonstrating any convincing generalisation is to have two observers interpret the data. However, they cannot be any two observers. To satisfy the requirement of reliability, they must show a certain kind of agreement in their interpretation of their observations; to satisfy the requirement of validity, they should also show that they are different in some relevant way. The question is whether or not it is possible to make explicit the way the observations and their interpretations should be equivalent and the way they might be "materially different" in order to enhance generalisability or validity.

An attempt will be made to suggest one possible way of having variation among observers, while retaining generality of interpretations among these observers, following the description of an example which prompted our concern with this question.

3. AN EXAMPLE

The example involves the coding of thirteen aspects of an interview with 40 boys and 40 girls aged between 9 years and 15 years. The interviews involved questioning the children about the following four aspects of their social world: community; government; social class; and school. The aim of these interviews was to classify children with respect to their level of social understanding within a framework of stage-sequence theories of cognitive development. The interview transcripts were classified by two observers into one of six levels. The interviews were not scored for accuracy of any content, but according to levels of understanding ranging from a pre-operational "playful" level to a systematic-analytical level.

Observer Tasks

The two expert judges who classified the interviews will, for consistency with the earlier general discussion, be called observers.

The topic and the order and number of the questions within each topic is shown on the right of Table 1. The other values shown in this table will be described in the next section. Preliminary to the coding of these 80 interviews, the observers discussed the system with the first author, and rehearsed their coding on pilot interviews.

For the main study reported here, the observers first coded the answers to the first two questions on community independently, and then compared their results. Wherever they disagreed, they discussed their classification, and reached a consensus. This procedure is common in the classification of data

according to stage theories of development. Following this work, they coded the 80 interviews on the five questions about government independently again, and then once again compared their results and reached a consensus where they disagreed. The same procedure was followed for the three questions on social class and for the four on the school.

The breaking up of the independent classifications and the discussion sessions across the 13 questions into four sets was not done with this study in mind. Instead, it was done in order to ensure consistency within observers across interviews with respect to the same sets. Classifying all 13 different questions from four sets appeared unduly taxing on the observers. In all, the classifications and discussions took each observer 40 hours of time. Thus the availability of these data is somewhat fortuitous.

Model of Analysis

The data were analysed according to a latent trait psychometric model which has been applied to ratings of attitude questionnaires constructed in the Likert tradition (Andrich, 1978) and to other performances involving ratings (Masters, 1982; Wright and Masters, 1982). The model belongs to the Rasch (1960, 1980) class of models. For the rating of the performance of a single person n with location parameter β_n on a single task or question i with location parameter δ_i , the model takes the form

$$P_{xni} = p(x_{ni}; \beta_n, \delta_i, \tau) = \frac{1}{\gamma_{ni}} \exp [x(\beta_n - \delta_i) - \sum_{k=1}^x \tau_{ki}] \quad (1)$$

where $x \in \{0, 1, \dots, m\}$ is the rating on the question, $\tau_{1i}, \tau_{2i}, \dots, \tau_{mi}$ are m cut-off points with respect to the adjacent ordered categories on the continuum for question i , and γ_{ni} is a normalising factor. The threshold parameters of Model (1) may be reparameterised in a way which make the new parameters somewhat more readily interpretable with respect to the distribution of the responses (Andrich, 1983).

The reparameterised model takes the form

$$p_{xni} = p(x_{ni}; \beta_n, \delta_i, \theta_i, \eta_i) =$$

$$\frac{1}{\gamma_{ni}} \exp [x(\beta_n - \delta_i) + x(m-x)\theta_i + x(m-x)(2x-m)\eta_i] \quad (2)$$

where now θ_i reflects the dispersion of the responses with respect to question i and η_i reflects the skewness. The dispersion parameter θ_i gives the average half distance between successive thresholds while η_i gives one sixth of the deviation from this average. The greater the value of θ_i , the greater the distance between thresholds, and this corresponds to a greater number of responses in the central categories. That is, the greater the value of θ_i , the smaller the dispersion of responses. Analogously, the greater the value of η_i in magnitude, the more the distribution is skewed, with more positive values indicating a wider spread of thresholds in the upper part of the continuum.

In this paper we concentrate on the location and dispersion parameters with respect to each question of the interview. However, for completeness, the probability distribution with respect to an asymmetrical set of thresholds is shown in Figure 1.

INSERT FIGURE 1 HERE

Stage Theories and Mathematical Models

Before proceeding with the interpretation of the analysis, some particularly pertinent points about the model and its relationship to stage theories of development are made. When translated into terms of a mathematical response model, stage theories of development in general, and of cognitive development in particular, are of the unfolding type. That is, when a person is in a given stage, he or she tends to respond in a way that

reflects that stage, and tends not to respond in ways which reflect either earlier stages or later stages. This is the case even though the stages are seen as hierarchical with the transition through each stage as both essential and invariant in reaching any particular stage. The usual cumulative models of traditional test theory and of Guttman scaling are essentially not appropriate.

The important feature of model (1) is that, with respect to the classification of persons into stages, it is of the unfolding type. Furthermore, it is probabilistic in nature, thus allowing for some random fluctuation which can be attributed to an inconsistency of a person's classification, due perhaps to a misunderstanding or lapse of concentration either by the respondent or the observer. With respect to this classification, there is presumed an underlying continuum on which the location parameter β_n may be mapped. Whether or not there is a real continuum of cognitive development is not at issue; the point is that it is convenient to conceive of such a continuum in classifying individuals. The greater the value of β_n , the more likely the person will be classified in a higher stage of development. The unfolding pattern is evident with respect to all categories except the first and the last, where the finite number of categories imposes constraints on classifications no matter how high or low the value of β_n .

The threshold parameters can be mapped on the continuum and they represent transition points in the classification. Thus if, for example,

$\beta_n = \tau_2$, then the probabilities that a person will be classified in stage 2 or stage 1 will be identical. (Notice that in the case of $m + 1$ stages, the first stage is scored as 0 and the final stage as m .) The location parameter δ_i of question i can also be mapped onto the same continuum. Thus even though a set of tasks may be designed to elicit

responses at the same level, it is possible that different questions facilitate classifications at different levels. That is, the same people may tend to be classified higher with respect to one question than with respect to another.

One final point to note regarding model (1) is that, because it is a member of the class of Rasch models, the total score of person n given by $r_n = \sum_i x_{ni}$ is sufficient for β_n . As a result, and to the degree that the model holds, to that degree the relationships among the task and threshold parameters are independent of the stages of development of the persons. Of course, the range of persons is important, and in this study, the range was substantial.

The classifications of the responses of the children were analysed in two ways. First, each classification by each observer was treated as a separate response, thus giving $2 \times 13 = 26$ separate task-observer combinations. Second, the consensus score from the two observers was treated as a single classification giving 13 task-observer combinations. The scale parameter estimates are examined first, followed by the location parameters.

4. RESULTS

Interpretation of Scale Parameters

Table 1 shows three columns of numbers. The first two columns show the values of the scale parameter for each of the two observers on each task. The third column shows the scale values for an analysis in which only the consensus scores were used.

INSERT TABLE 1 HERE

The first observation to make is that, except for two of the 13 questions, the scale value for the consensus score is greater than either of the two scale value based on the individual scores. This result, which implies that the consensus classification tended to be more central than the individual classifications, is perhaps not surprising, in that the very idea of consensus suggests some ground which is most common to the two graders.

The second observation, and perhaps a little more surprising, is that the scale values increase systematically for both observers and for the consensus score from the first set of interviews to the last.

It may be argued that some variation in the responses is due to the nature of the questions. However, the pattern is so strong, that it is difficult not to conclude that the discussions in reaching a consensus on one set of interviews affected the subsequent classifications, and that it affected them in such a way that, once again, more and more central scores were assigned. This effect of regression is manifested vividly between the second and third classification sessions, and the third and fourth sessions.

Clearly the effects of both types of regression is to reduce the discrimination among the persons classified.

Interpretation of Location Parameters

Having identified an effect attributed to reaching a consensus in the process of coding the interviews, it is instructive to consider whether there is any observable effect on the location parameters. Table 2 shows these values for the analysis with 26 separate classifications and the 13 consensus classifications. In addition, the differences between the location values for the two observers with respect to the 13 tasks are shown.

INSERT TABLE 2 HERE

A pattern in the order of classification is again discernable. In the first classification, observer J had lower values than observer R, indicating that, because she tended to classify interviews into higher categories, she was more lenient. In contrast, in all the second set of classifications, observer R is the more lenient. The relative leniency is then reversed once again for the third set. In the final set, the differences between the two observers are generally smaller in magnitude, with two of the three showing observer J more lenient and one showing observer R more lenient. Again, it is hard not to conclude that the discussions between the observers with respect to one set of classification affected the subsequent classification. In particular, it seems that in attempting to accommodate each other's position, the lenient observer became relatively harsher and the harsher observer became more lenient after the first two discussion sessions.

Contingency Table Interpretation

The use of the psychometric model of equation (2) displayed clearly the effects across the rating sessions. The actual effect on the classifications can be appreciated more directly with contingency tables with respect to each question. It suffices to show such tables for only three questions, and Table 3 displays these for questions 1, 7 and 13, that is, the first question, the middle question and the last question.

INSERT TABLE 3 HERE

These tables show the trend identified earlier. Table 3a has the largest spread of scores across the whole range, Table 3b the next largest range, and Table 3c the smallest range. Most interesting however, is that Tables 3a and 3b show a percentage of classifications with perfect agreement of only 51 per

cent and 52 per cent, while Table 3c shows the much higher agreement rate of 85 per cent. Clearly, a superficial examination of percentage agreement would suggest that the final question was classified more reliably than either question 1 or question 7 and that the effect of the discussions was producing a desirable result. On the other hand, the discrimination among the classifications is much lower in the last question, and it seems unavoidable to ask whether or not the classification of the final question is as valid as is that of the earlier questions. It is possible to argue that, because the observers were discriminating less, that the classifications were less valid.

The concern can be made even more vividly by considering the implications of a set of scores all of which had the same distribution as the last question in which only one out of 80 children was given a score outside the range 2 to 4. In such a case, the range of classification of the children would be substantially reduced, and no child would be identified as being in either of the first two stages of development and very few would be identified as being in the highest stage.

5. DISCUSSION

The above example illustrated that, as the two observers compared and discussed their classifications, they tended to influence each other in their subsequent classifications. This influence was such that they both tended to assign more central or more conservative scores. These more conservative scores assigned later in the process showed greater agreement than those assigned earlier. The question posed is whether the scores showing greater agreement were more valid than those showing less agreement.

Clearly, the results of the discussions between the observers did not necessarily have to create the effect identified above. The scores might not have regressed at all. That they did regress is an empirical observation which need not always follow. However, that the effect may generalise to other situations does not seem unreasonable.

The above example also makes clear that the observers are an integral part of the data and that their differences, and their similarities, cannot simply be taken for granted. Further, if we argue that the scores with greater agreement, but with less variability, are possibly less valid than the scores with greater variability and less agreement, we need to have a procedure which can deal with such data.

We suggest that the models of equations (1) and (2), which helped reveal the effect we identified, can also be used to account for the differences among observers. By treating each task-observer classification as a separate response, the model accounts for differences in leniency among task-observer combinations, differences in dispersion for the same combinations, as well as differences in skewness. If the disagreements in classification are due to systematic differences in these three factors, then the use of the model accounts for any differences in disagreements.

It should be stressed that the use of the model does not completely obviate the need for observers to be trained and to compare their classifications. They need to go through these processes to the point where their classifications are conformable as evidenced by conforming to the suggested model; however classifications need not be identical.

We are now in a position to return to the main question posed in Section 2: is it possible to make explicit the way the observations should be equivalent between observers and the way they might be 'materially different' at the same time?

We suggest that observers who have a difference in leniency, and perhaps a difference in spread of their classifications, are 'materially different'. On the other hand, if these differences are systematic and can be accounted for by a model, then they are also equivalent for the purpose of any inferences. The model described above is one of a number available from person-item response theory (PIRT), and therefore the great deal of literature

available in this field may be invoked to understand and deal with issues of inter-observer agreement.

The case presented above leads to the conclusion that enhancing reliability among observers as defined traditionally may reach a point where further increases in reliability reduce validity. However, this trade-off between reliability and validity is not unusual, and the particular case which is analogous to the one described above is the so called attenuation paradox of traditional test theory. An example of the paradox is that while the responses to two items of a test need to be correlated in order to reveal that they test the same abilities, the correlation may become so high that the second item is redundant. Consequently, the validity is not enhanced by the use of the two items instead of just one of them. The parallel between the case of two graders and two items is obvious.

The relevant question to consider here is when is the correlation between two graders "too high". The answer may be also available from PIRT. Andrich and Pedler (1983) argue that the attenuation paradox arises when the universal assumption of local independence in PIRT is violated. This assumption is violated when, for individuals of a fixed ability, responses to items are nevertheless correlated. Analogously, then, if for fixed abilities or levels, the classifications among observers are correlated, then the reliability among the observers is "too high". How evidence that classification for persons of the same level or ability are correlated may be obtained will not be explored here, but it is evident that once again considerations from PIRT may be used to help understand issues of inter-observer agreement.

REFERENCES

- Andrich, D. and Hake, R. Two programs for analysing essay writing and grading performances, Chicago State University, 1979.
- Andrich, D. A rating formulation for ordered response categories. Psychometrika, 1978, 43, 561-573.
- Andrich, D. Diagnosing and accounting for response sets provoked by items of a questionnaire. National Council of Measurement in Education, Montreal, 1983.
- Andrich, D. and Pedler, P. The attenuation paradox in latent trait theory. Symposium on Item Response Theory, National Council of Measurement in Education, Montreal, 1983.
- Barnes, B. T. S. Kuhn and Social Science. London, Macmillan, 1982.
- Dawes, R.M. Fundamentals of Attitude Measurement. New York, John Wiley, 1972.
- Guilford J.P. Psychometric Methods. New York, McGraw-Hill, 1954 (2nd Edition), McGraw Hill, 1936 (1st Edition)
- Keating, D.P. and Clark, L.V. Development of Physical and Social reasoning in adolescence. Developmental Psychology, 16, 1980, 23-30.
- Kuhn, T.S. The structure of scientific revolutions. University of Chicago Press, Chicago, 1970.
- Masterman, M. The nature of a paradigm. In Criticism and the growth of knowledge, I. Lakatos and A. Musgrave (Ed) Cambridge, Cambridge University Press, 1970.
- Masters, G.N. A Rasch model for partial credit scoring. Psychometrika, 1982, 47, 149-174.
- Rasch, G. Probabilistic models for some intelligence and attainment tests. (Copenhagen, Danish Institute for Educational Research, 1960), expanded edition, Chicago, The University of Chicago Press, 1980.
- Stanley, J. Analysis of unreplicated three-way classifications, with applications to rater bias and trait independence. Psychometrika, 1961, 26, 205-219.
- Wright, B.D. and Masters, G.N. Rating Scale Analysis, Chicago, MESA PRESS, 1982.

TABLE 1

Scale values for classifications of two observers
treated separately and for a consensus score on thirteen questions

Question	Observer J		Observer R		Consensus C		
	(Mean)*		(Mean)*		(Mean)*		
1	.913		.887		1.073		Set 1 Community
2	1.021	.97	1.081	.98	1.163	1.118	
3	.973		1.243		1.088		Set 2 Government
4	.938		.924		1.070		
5	.957	1.18	1.309	1.11	1.103	1.343	
6	1.623		1.160		1.818		
7	1.414		0.919		1.638		
8	1.547		1.323		1.849		Set 3 Social Class
9	1.879	1.74	1.148	1.26	1.910	2.098	
10	1.802		1.318		2.535		
11	2.065		1.540		2.119		Set 4 School
12	2.322	2.09	1.359	1.54	2.111	2.098	
13	1.892		1.719		2.064		

* Mean of scale values per set.

TABLE 2

Location values for classifications of two observers
treated separately and for a consensus score on thirteen questions

Question	Observer J	Observer K	Difference	Consensus	
1	.546	1.369	.823	.814	Set 1 Community
2	.304	1.278	.974	.401	
3	.109	-0.597	-.787	.260	Set 2 Government
4	.362	-0.345	-.707	.425	
5	.479	-0.271	-.750	.630	
6	1.831	0.479	-1.352	1.989	
7	0.094	-0.190	-0.380	.082	
8	-1.093	0.184	1.277	-0.912	Set 3 Social Class
9	-0.769	0.834	1.603	-0.324	
10	-1.772	1.282	3.054	-2.239	
11	-0.692	-0.805	-0.108	-0.818	Set 4 School
12	-0.696	-0.284	0.412	0.378	
13	0.565	-1.148	-0.583	-0.687	

TABLE 3

Distribution of classifications by two observers
on three questions

		Observer R					
		0	1	2	3	4	5
Observer J	0	2(2)*					
	1	5	4(11)	2			
	2		7	7(15)	4	1	
	3			3	16(27)	4	
	4			4	5	8(18)	
	5		1		1	2	2(5)

per cent
agreement:
51

3a : Question 1

		Observer R					
		0	1	2	3	4	5
Observer J	0						
	1		3(3)				
	2		6	7(21)	8		
	3			3	21(35)	8	1
	4			1	2	9(18)	9
	5						1(2)

per cent
agreement:
52

3b : Question 7

		Observer R					
		0	1	2	3	4	5
Observer J	0						
	1						
	2			36(37)	8		
	3			1	18(26)	2	
	4					14(16)	1
	5						1

per cent
agreement:
85

3c : Question 13

* Bracketed scores represent the consensus distribution.

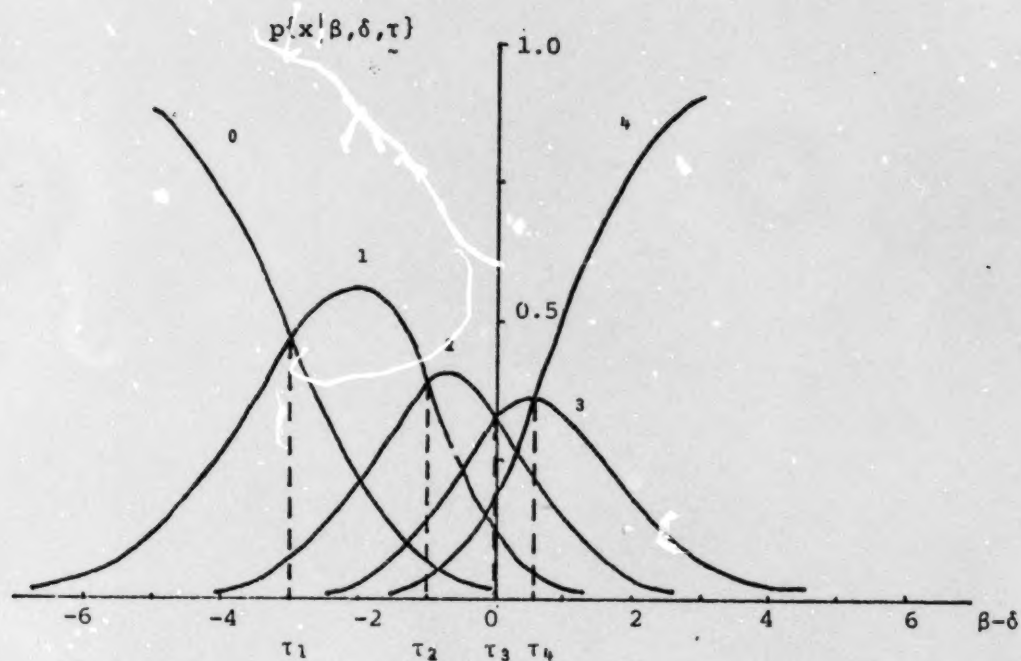


Figure 1. Probabilities of classification in stage x as a function of the person's location, the task's difficulty and the location of the boundaries between stages.